# A Novel Thresholding for Prediction Analytics with Machine Learning Techniques

**Shakir Khan** [1*, 2], **Reemiah Muneer Alotaibi** [1]

sgkhan@imamu.edu.sa

[1] College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

[2] University Centre for Research and Development, Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, India

Corresponding Author: Shakir Khan* (sgkhan@imamu.edu.sa )

## Summary

Machine-learning techniques are discovering effective performance on data analytics. Classification and regression are supported for prediction on different kinds of data. There are various breeds of classification techniques are using based on nature of data. Threshold determination is essential to making better model for unlabelled data. In this paper, threshold value applied as range, based on min-max normalization technique for creating labels and multiclass classification performed on rainfall data. Binary classification is applied on autism data and classification techniques applied on child abuse data. Performance of each technique analysed with the evaluation metrics.

***Keywords:***

*Machine learning, binary classification, multiclass classification, multi-label classification, threshold, cluster-then-predict.*

## 1. Introduction

### 1.1. Data Analytics

Analytics of data is sketching some conclusions about data. Data analytics is a component of data science, which is combined areas as mathematics, statistics and computer science. Statistical methods are well suitable for solving many of the problems. There are dissimilar kinds of analytics methods like descriptive, predictive, diagnostic and prescriptive. Data analytics have a potential to articulate a fruitful solution for experimentation [1].

### 1.2. Predictive Analytics

Machine learning is stem of artificial intelligence which occupies a significant position in computational intelligence. Every domain such as medical, agriculture, Finance are using machine learning to fulfil the needs of their decisive state of affairs. There are supervised and unsupervised algorithms are using based upon the data. Analysis of preceding data provided useful future resolutions are derived is known as predictive analytics. Classification algorithms are effectively working for prediction. Predictive analytics can be applied on different kinds of data such as qualitative and quantitative. Regression is another technique to prediction especially for continuous data. Prediction in regression is performs with the dependent and independent variables [2,3,4,5].

### 1.3. Binary Classification

Autism data consisting of two labels named as autism, which is denoted label as “YES”, and non-autism which is denoted label as “NO”. Binary classification is one the machine learning technique which is handling two class labels. ASD is an emerging social problem in childhood of human. In 1970, autism society of United States arranged an awareness program to light a confident on autism people. Autism awareness month as April was announced by autism society of United States. World health association (WHO) reveals that 1 among 160 are affected because of autism [IAP]. India is having 1 among 500 individuals .Autism is identified above two years clearly in human life. Mostly boys are affecting comparing with the girls and by this psychiatric disorder, parents getting more saddle with economical and mental depression for care concerning the children. Medical Analysis exposed that autism disorder splits as hypersensitivity and hyposensitivity by causing confusion among the children [6, 7, and 8.

### 1.4. Multi-class classification

Classification of more than two instances or classes is called as multi-class classification, which is also known as multi-nominal classification. There are plenty of

Manuscript received January 5, 2023
Manuscript revised January 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.1.5

algorithms are available for classification technique. Most suitable algorithms for multi class classification are KNN, Decision trees, Naïve Bayes, Random forest, gradient boosting, SVM, C4.5, Neural network, ID3 [8, 9, 10, 11] (Baidaa M, 2019).

### 1.5. Multi-label classification

Most of the research has dealt with single label classification. Multi-label classification is a method of handling more than two labels for a class. There are a couple of methods are available in this category. This can be divide into problematic transformation approaches which is transform multi label to single label classification technique and another one is algorithm adaptation methods. KNN, decision trees and neural networks are most appropriate algorithms for multi-label classification [12] (Raed Alazaidah, 2015).

### 1.6. Threshold value

Threshold is a probability point which is limits the criteria for doing certain process on data such as classification and clustering and association rules. Threshold value will smooth data. Largest value of threshold will neglect more coefficients then over smoothing occur on result. On the other side, smallest threshold will take more coefficients, and then the resultant data get poor performance [13]. (K. Sasirekha, 2014).

## 2. Background of Study

The research of autism reveals that there is 1 among 88 has been affected in United States. Comparing with children, adults are distressing by autism spectrum disorder worldwide. Intensifying of this disability will be reach 2 million persons with ASD in US [14, 15].

Classification techniques of data mining like neural network (NN), support vector machine (SVM) and proposed fuzzy logic IF THEN rules was taken for classify autism as mild moderate and severe. SVM provided enhanced performance than the other techniques [16]. Regression is a machine learning technique which knob with binary outcome. Screening methods consists of questionnaires from A1 to A10 in the ADI (Autistic Diagnostic Interview). ADI will be the training and many of the questions asked to individuals with the help of mobile application ASD Tests. Filtering methods Information Gain (IG) and Chi Square Testing (CHI) were applied on the features of Adolescent data. Logistic regression was the fit classifier for binary outcomes [17, 18]. [19, 20] analysed ASD diagnosis. Logistic regression applied as a fit classifier and then performance metrics such as accuracy, F1 score and recall were used to analyse the feasibility of the model. Comparison of ML algorithms like Support vector machine, logistic regression, naive bayes, K-nearest neighbour and CNN have been applied on ASD data. Among the performance of these algorithms SVM supported well for classifying ASD data. Statistical analysis of ANOVA has been used to show the high ability of regional synchronization likelihood with low frequency bands. Random forest and decision trees are the easy way of implementation and tree form of result exhibited among other algorithms of classification [21].

Forecasting of landslides in Kalimpong Region of the Darjiling Himalayas hills were illustrated with threshold of rainfall level. Cumulative rainfall and duration of rainfall taken as components and power law equation has been used to set the threshold value for identify the landslide [22]. Prediction of rainfall estimated with data of climatic conditions such as temperature of wind speed, pressure, dewpoint, humidity. Decision tree applied on this data and gini index used for better accuracy [23]. [24] was used machine learning techniques like bayesian linear regression, boosted decision tree regression, auto correlation function, decision forest regression were applied on rainfall data of Terengganu of Malaysia.

.

Half of the children among world-wide facing violence. So, consideration is needed for the violence against children.AI helps to overcome this problem in numerous ways. Another approach is mHealth, which will be in the part of prevention of child abusing [25, 26]. Child abusing is everywhere in worldwide now a days. Public health institution of Netherlands country tried to build a decision support system to prevent abusing against children. Using unstructured data of child abusing, text mining methods has been applied and decision support API developed [27]. Online media are one of the causes which are affecting society. Text mining and feature extraction done on child abuse data and classification algorithms were applied [28].

Child maltreatment data handled with big data intelligent techniques like c4.5 algorithm and apriori algorithm which used for identifying changes and trends [29]. Resultant clusters were stored in HIPPO cluster.

Another data mining technique nearest neighbor used to predict of risk among children in society [30]. A sentiment analysis of customers regarding the products was done on social media data with k-means algorithm and CART algorithm. Classification process provided fruitful prediction results [31, 32].

Gene expression of patient responses regarding the drug tamoxifen' data were taken for experiment. T-time threshold, which is a distance threshold value between two time series data. Structures of cluster were affected when the threshold value changes [31, 32]. Stationary wavelet transform was taken for the threshold value for removing Gaussian noise of an image of fingerprint.

## 3. Methodology

### 3.1. Flow of the process

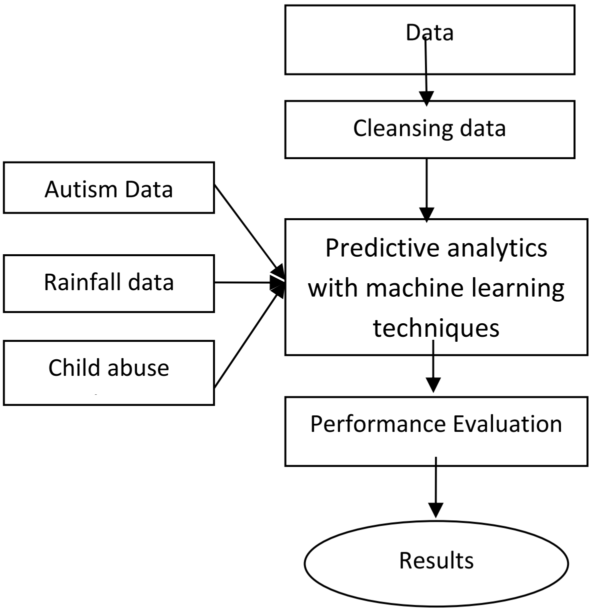

Figure.1: Flow of the process

### 3.2. Depiction of data

Autism child data is having 21 attributes and 292 instances. There are information regarding the children such as age, gender, ethnicity, born with jaundice, family member with PDD, who is completing the test, country of residence, used the screening app before, screening method type and ten questions of screening method which consists of yes or no answers. This data downloaded from UCI repository. Rainfall data consists of 16 attributes and 115 instances. This data is used from government of India website. Child abuse data consists of 14 attributes and 13 instances which is used from NCRB.

### 3.3. Cleansing Data

Feature selection is one of a technique to select most important features of data. Principal component analysis (PCA) is feasible method for feature selection, which was applied in autism data. In this paper, missing values handled for all the three data.

### 3.4. Predictive Analytics

Proposed Technique for threshold determination

Our proposed method for setting threshold is based on min-max normalization technique.

$$a' = \frac{a - \min(a)}{\max(a) - \min(a)} \qquad \text{--------------- (1)}$$

Min-max normalization technique is an effective technique to normalize data and modernize into 0 to 1 without affecting the originality of data.

Threshold setting for multi class classification

```
Input: statistical report of data
output: class label
  • Load data
  • Select class attribute $a_i$.
  • Apply min-max normalization technique to dataset D.
  • If $a_i<0.33$ then
        Label "LOW"
    If $0.34 \leq a_i \leq 0.66$ then
        Label"MEDIUM"
    If $0.67 \leq a_i \leq 1.0$ then
        Label"HIGH"
  • Apply labels to $a_i$.
```

### 3.5. Cluster-then-predict:

Cluster-then-predict is a methodology for unlabelled data. Basically clustering technique applied to the data and then classification will be done through the clusters .Resulting clusters are taking as features for the classification process.

## 4. Results and Discussion

### 4.1. Pre-processing

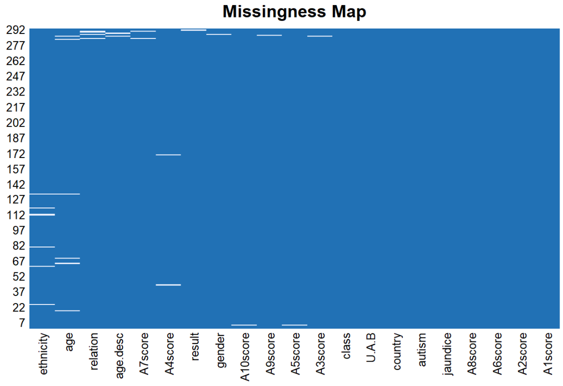

Figure.2: Autism data with missing values

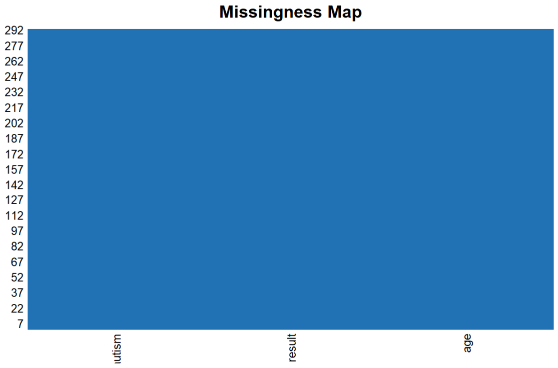

Figure.3: Autism data without missing values

The above picture shows that autism data contained missing values in the attributes such as ethnicity, age and A1score.Missing values were filled with mean value of the respective columns. Figure 2&3 are missmap of a data.

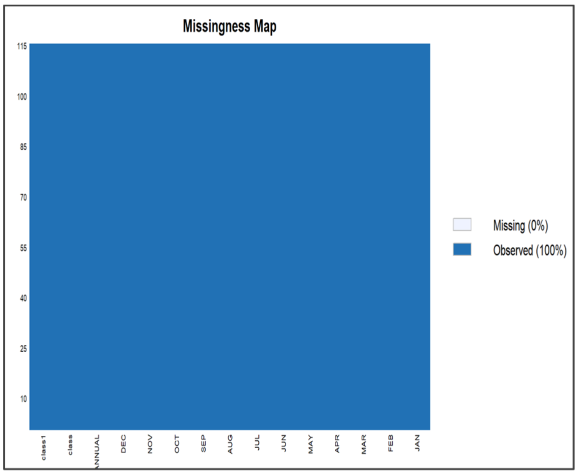

Figure.4: Missmap of rainfall data

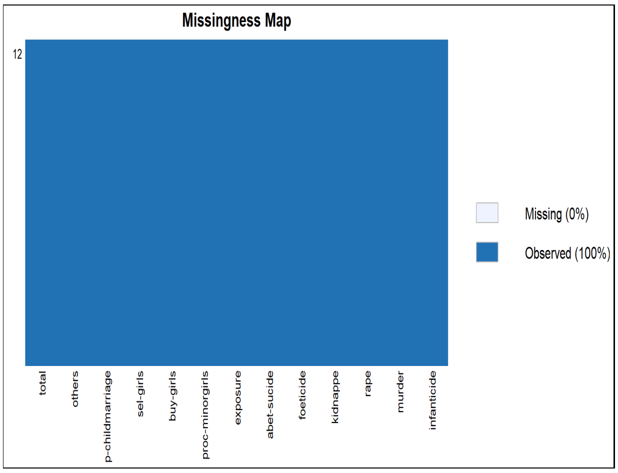

Figure.5: Missmap of child abuse data

### 4.2. Predictive Analytics

Evaluation Metrics of Autism data:

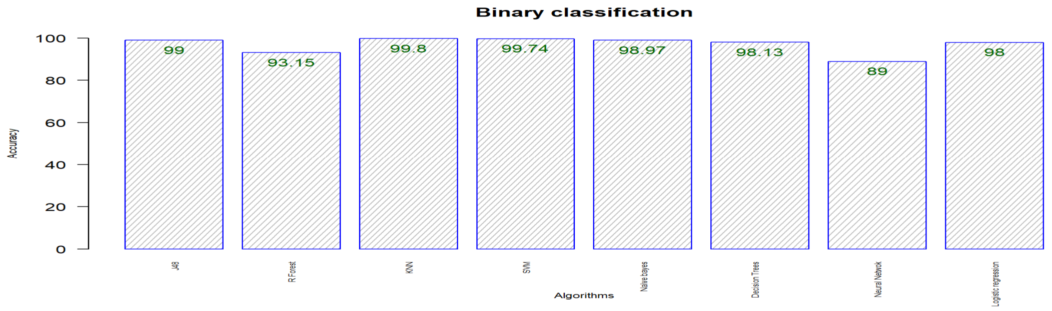

Figure.6: Comparison of accuracy for ASD data

| Algorithm | Precision | Recall | F1 Score | MAE | MSE | RMSE | TPR | FPR |
|---|---|---|---|---|---|---|---|---|
| J48 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| Random Forest | 0.95 | 0.9 | 0.93 | 0.1 | 0.01 | 0.13 | 0.93 | 0.067 |
| KNN | 1 | 1 | 1 | 0.11 | 0.10 | 0.33 | 0.88 | 0.11 |
| SVM | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| Naïve bayes | 1 | 1 | 1 | 0.04 | 0.01 | 0.1 | 0.99 | 0.01 |
| Decision Trees | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| Neural Netwok | 1 | 1 | 1 | 0 | 0.09 | 0.31 | 1 | 0 |
| Logistic regression | 1 | 1 | 1 | 0.13 | 0.13 | 0.36 | 1 | 0 |

Table.1: Evaluation metrics for ASD Data

J48, KNN, SVM, naïve bayes, decision trees, neural network, and logistic regression are the classification and regression techniques are preserved for binary classification. KNN provides 99.8% of accuracy among the other algorithms for autism data.Table.1 explores error metrics such as MAE, MSE, and RMSE are near to 0 and quality metrics such as precision, recall and f1 score are near to 1.so, the above algorithms performed well on autism data.

### 4.3. Evaluation metrics of Rainfall data

Min-max technique on Rainfall Data:

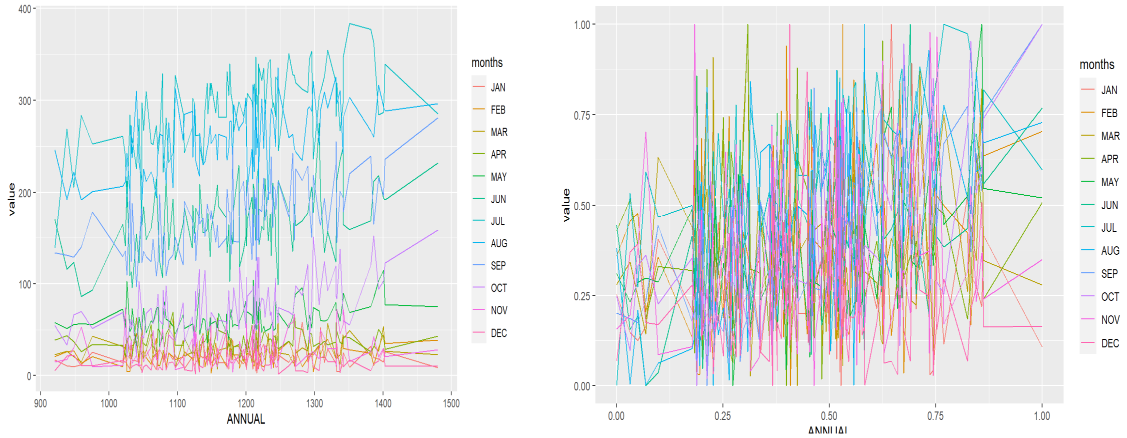

Figure.7: Rainfall data before and after Min-Max technique

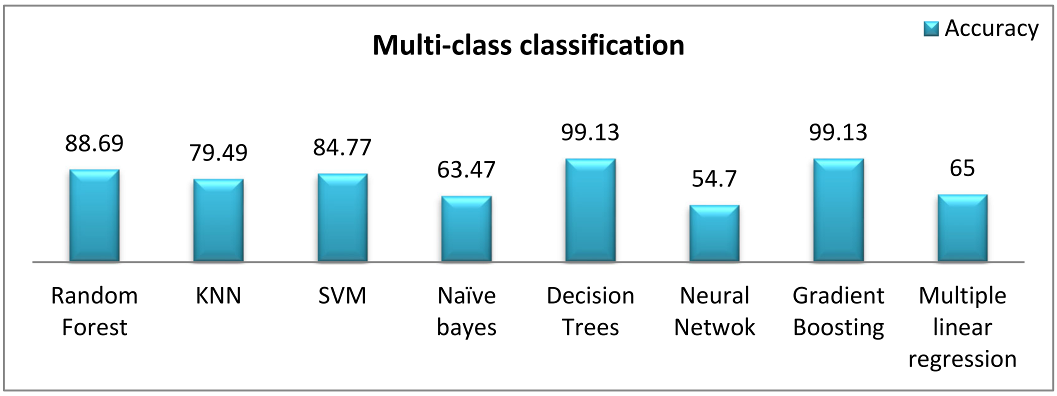

Figure.8: Accuracy for rainfall data

Figure.8 shows that the accuracy of different kinds of algorithms for rainfall data which is labelled using min-max threshold technique.Table.2 displays evaluation metrics and error mostly near to 0 and then the quality metrics mostly near to 1.Neural network produced lowest accuracy comparing with other algorithms.

| **Algorithm** | **Precision** | **Recall** | **F1 Score** | **MAE** | **MSE** | **RMSE** | **TPR** | **FPR** |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 0.99 | 0.99 | 0.99 | 0.06 | 0.133 | 0.017 | 0.95 | 0.002 |
| KNN | 0.80 | 0.88 | 0.84 | 0.02 | 0.01 | 0.13 | 0.97 | 0.032 |
| SVM | 1 | 1 | 1 | 0.2 | 0.07 | 0.27 | 1 | 0 |
| Naïve bayes | 0.35 | 0.52 | 0.54 | 0.2 | 0.12 | 0.36 | 0.63 | 0.44 |
| Decision Trees | 0.99 | 0.99 | 0.99 | 0.007 | 0.004 | 0.07 | 0.99 | 0.003 |
| Neural Netwok | 0.50 | 0.33 | 0.32 | 0.3 | 0.1 | 0.4 | 0.64 | 0.23 |
| Gradient Boosting | 1 | 1 | 1 | 0.13 | 0.13 | 0.36 | 1 | 0 |
| Multiple Linear Regression | 0.54 | 0.37 | 0.76 | 0.3 | 0.1 | 0.36 | 0.69 | 0.33 |

Table.2: Evaluation metrics for rainfall data

### 4.4. Evaluation metrics for child abuse data

Child abuse data consists of numerical data without class labels. Based on cluster-then-predict method, prediction can be done linear regression, naive bayes and decision tree algorithms are most supported than the other classification algorithms for child abuse data. Bayesian regression is another pretty method for prediction analytics, which shows higher accuracy than the other techniques.

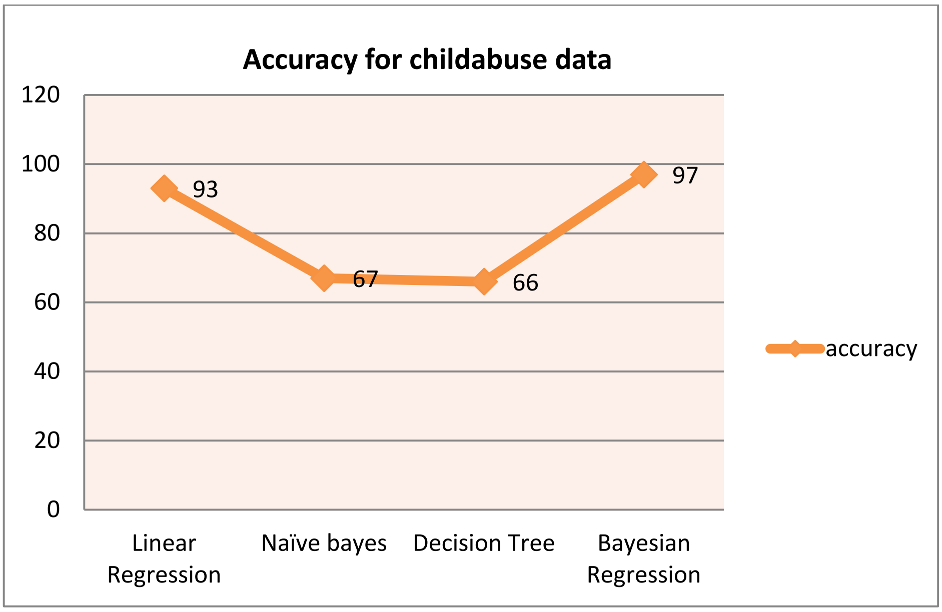

Figure.9: Child abuse data results

## 5. Conclusion

In this paper, there are different kinds of data were taken for prediction analytics. Binary classification techniques applied on autism data which consists of 0 and 1 data. Rainfall data having continuous data and multi-class classification techniques applied on this data. Child abuse data contains numerical data cluster-then-predict methodology and bayesian regression has applied on this data. Other kinds of analytical methods can be applied for these data in the future era.

## References

1. S. Khan, "Data Visualization to Explore the Countries Dataset for Pattern Creation," *International Journal of Online Biomedical Engineering,* vol. 17, no. 13, pp. 4-19, 2021.
2. Vaibhav Kumar, M. L. Garg, ,"Predictive Analytics: A Review of Trends and Techniques", Volume 182 – No.1, July 2018.
3. S. Khan, "Visual Data Analysis and Simulation Prediction for COVID-19 in Saudi Arabia Using SEIR Prediction Model," *International Journal of Online Biomedical Engineering,* vol. 17, no. 8, 2021.
1. P. Nikolaidis, M. Ismail, L. Shuib, S. Khan, and G. Dhiman, "Predicting Student Attrition in Higher Education through the Determinants of Learning Progress: A Structural Equation Modelling Approach," *Sustainability,* vol. 14, no. 20, p. 13584, 2022.
2. M. M. Akhtar, A. S. Zamani, S. Khan, A. S. A. Shatat, S. Dilshad, and F. Samdani, "Stock market prediction based on statistical data using machine learning algorithms," *Journal of King Saud University-Science,* vol. 34, no. 4, p. 101940, 2022.
3. Eunice Kennedy Shriver ,Centers for Disease Control and Preventionhttp://www.cdc.gov/ncbddd/autism National Institute,2019.
4. A. u. Haq, J. P. Li, S. Khan, M. A. Alshara, R. M. Alotaibi, and C. B. Mawuli, "DACBT: deep learning approach for classification of brain tumors using MRI data in IoT healthcare environment," *Scientific Reports,* vol. 12, no. 1, p. 15331, 2022/09/12 2022.
5. A. U. Haq *et al.*, "IIMFCBM: Intelligent Integrated Model for Feature Extraction and Classification of Brain Tumors Using MRI Clinical Imaging Data in IoT-Healthcare," *IEEE Journal of Biomedical Health Informatics,* vol. 26, no. 10, pp. 5004 - 5012, 2022.
6. A. K. Singh, I. R. Khan, S. Khan, K. Pant, S. Debnath, and S. Miah, "Multichannel CNN model for biomedical entity reorganization," *BioMed Research International,* vol. 2022, no. Article ID 5765629, p. 11 pages, 2022.
7. Baidaa M Alsafy, Zahoor M. Aydam, Wamidh K. Mutlag ,"Multiclass Classification Methods: A Review", ,International Journal of Advanced Engineering Technology and Innovative Science (IJAETIS),Volume 5, Issue 3, Page No: 01-10,2019,ISSN:2455-1651.
8. S. Khan, "Study Factors for Student Performance Applying Data Mining Regression Model Approach," *International Journal of Computer Science Network Security,* vol. 21, no. 2, pp. 188-192, 2021.
9. Raed Alazaidah , Fadi Thabtah , Qasem Al-Radaideh A Multi-Label Classification Approach Based on Correlations Among Labels, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015.
10. K. Sasirekha , K. Thangavel ,A Novel Wavelet based Thresholding for Denoising Fingerprint Image, 978-1-4799-5748-4/14/$31.00 © 2014 IEEE, International Conference on Electronics, Communication and Computational Engineering (ICECCE).
11. David C. Wyld , Tingyan Deng ,"Classifying autism spectrum disorder using machine learning models ", , DOI: 10.5121/csit.2021.110306, 2021.
12. S. Khan and M. Alshara, "Fuzzy Data Mining Utilization to Classify Kids with Autism," *International Journal of Computer Science Network Security,* vol. 19, no. 2, pp. 147-154, 2019.
13. M. S. Mythili, A. R. Mohamed Shanavas."A Study on Autism Spectrum Disorders using classification Techniques", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-4 Issue-5, Nov-2014.
14. Fadi Thabtah, Neda Abdelhamid and David Peebles,, Thabtah et al ,"A machine learning autism classification based on logistic regression analysis",. Health Information Science and Systems (2019) 7:12.
15. Raed Alazaidah , Fadi Thabtah , Qasem Al-Radaideh, "A Multi-Label Classification Approach Based on Correlations Among Labels", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015.
16. Daniel Bone, Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and promises, Journal of Autism and Developmental Disorders,
17. Brian McNamara, Camila Lora, Donyoung Yang, Fabiana Flores, Paul Daly, "Machine Learning Classification of Adults with Autism Spectrum Disorder", April 29, 2018.
18. Dr. R. Uma Rani, R. Suguna and Miss. P. Amsini, "A Study of Autism Spectrum Disorder Using Principal Component Analysis and Fuzzyc-means clustering", IJMA- 9(3), March-2018,
19. Togaru Surya Teja, Abhirup Dikshit and Neelima Satyam, Determination of Rainfall Thresholds for Landslide Prediction Using an Algorithm-Based Approach: Case Study in the Darjeeling Himalayas, India, MDPI, 2019.
20. Ayisha Siddiqua L, Senthil kumar N C,,"Heavy Rainfall Prediction using Gini Index in Decision Tree , International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

21. Rainfall forecasting model using machine learning methods: Case study terengganu, Malaysia,Wanie M. Ridwan , Michelle Sapitang , Awatif Aziz , Khairul Faizal Kushiar , Ali Najah Ahmed ,Ahmed El-Shafie, Ain Shams Engineering Journal 12 (2021) 1651–1663.
22. Xanthe Hunt, Mark Tomlinson, Siham Sikander , Sarah Skeen1, Marguerite Marlow1, Stefani du Toit 1 and Manuel Eisner Artificial Intelligence, Big Data, and mHealth: The Frontiers of the Prevention of Violence Against Children, 4,Frontiers in artificial intelligence,2020.
23. S. Khan and M. F. AlAjmi, "Impact of medical technology on expansion in healthcare expenses," International Journal of Advanced Computer Science Applications, vol. 4, no. 4, 2013.
24. Chintan Amrit, Tim Paauw, Robin Aly, Miha Lavric," Using text mining and machine learning for detection of child abuse, Computers and Society", https://doi.org/10.48550/arXiv.1611.03660,2016.
25. Xanthe Hunt, Mark Tomlinson, Siham Sikander , Sarah Skeen1, Marguerite Marlow1, Stefani du Toit 1 and Manuel Eisner Artificial Intelligence, Big Data, and mHealth: The Frontiers of the Prevention of Violence Against Children, 4,Frontiers in artificial intelligence,2020.
26. Mohammadreza Keyvanpour,Mohammadreza Ebrahimi,Necmiye Genc Nayebi ,Olga Ormandjieva and Ching Y. Suen," Automated Identification of Child Abuse in Chat Rooms by Using Data Mining", Data Mining Trends and Applications in Criminal Science and Investigations", DOI: 10.4018/978-1-5225-0463-4.ch009
27. Abdurazzag A Aburas, Mohammad Hassan, Hilary Lin, Shreshtha Batshu,"Child maltreatment forecast using Bigdata intelligent approaches",2018 ,Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS).
28. James Little1, Hayder A. Waheed and Andy Rixon," Evaluation of Data Mining for Two Child-related", 2018.
29. Rishabh Soni, K. James Mathai,An Innovative 'Cluster-then-Predict' Approach for Improved Sentiment Prediction,Advanced Computing and Communication Technologies (pp.131-140), DOI:10.1007/978-981-10-1023-1_13
30. M. AlAjmi and S. Khan, "PART OF AJAX AND OPENAJAX IN CUTTING EDGE RICH APPLICATION ADVANCEMENT FOR E-LEARNING," in *INTED2015 Proceedings*, 2015, pp. 4058-4063: IATED.
31. Johannes Aßfalg, Hans-Peter Kriegel, Peer Kr¨oger, Peter Kunath, Alexey Pryakhin, Matthias Renz, T-Time: Threshold-Based Data Mining on Time Series, In Proc. 24th International conference on data engineering (ICDE'08),Mexico 2008.
32. S. Khan and H. Alghulaiakh, "ARIMA Model for Accurate Time Series Stocks Forecasting," *International Journal of Advanced Computer Science Applications,* vol. 11, no. 7, pp. 524-528, 2020.
33. https://towardsdatascience.com/cluster-then-predict-for-classification-tasks-142fdfdc87d6/cole-feb-11.2020
34. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.
35. https://iapindia.org/pdf/child-india/2021/CHILD-INDIA-APRIL-2021.pdf